Rakamin Academy X KALBE Nutritionals

Submission Final Task

Kalbe Nutritionals Data Scientist Virtual Internship Program

Presented by
Marselius Agus Dhion

Marselius Agus Dhion

About You

Saya merupakan mahasiswa semester 5 di Universitas Kristen Maranatha dengan jurusan Sistem Informasi. Saya memiliki *interest* di bidang data analyst dan data science. Oleh karena itu, saya memiliki keahlian menggunakan SQL (MySQL, PostgreSQL, Server SQL), Python, dan tools visualisasi seperti Tableau dan Looker Studio.

| LinkedIn | Curriculum Vitae | Github | Tableau | Medium |

Rakamin Academy

Job Experiences

Data Analyst Intern – PT Kimia Farma
(November 2022 - January 2023)

Saya menganalisis total revenue secara keseluruhan, revenue dari cabangnya, dan total produk yang terjual dari suatu periode menggunakan Looker Studio.

Data Analyst Intern – Ditusi Gaming
(June 2023 - July 2023)

Saya menganalisis konten dan engagement social media Tiktok dan Instagram. Seperti menganalisis konten yang diminati dengan metrics likes, share, new followers, Engagement Rate Post (ERP), Engagement Rate Reach (ERR), dan beberapa metrics lainnya menggunakan Looker Studio.

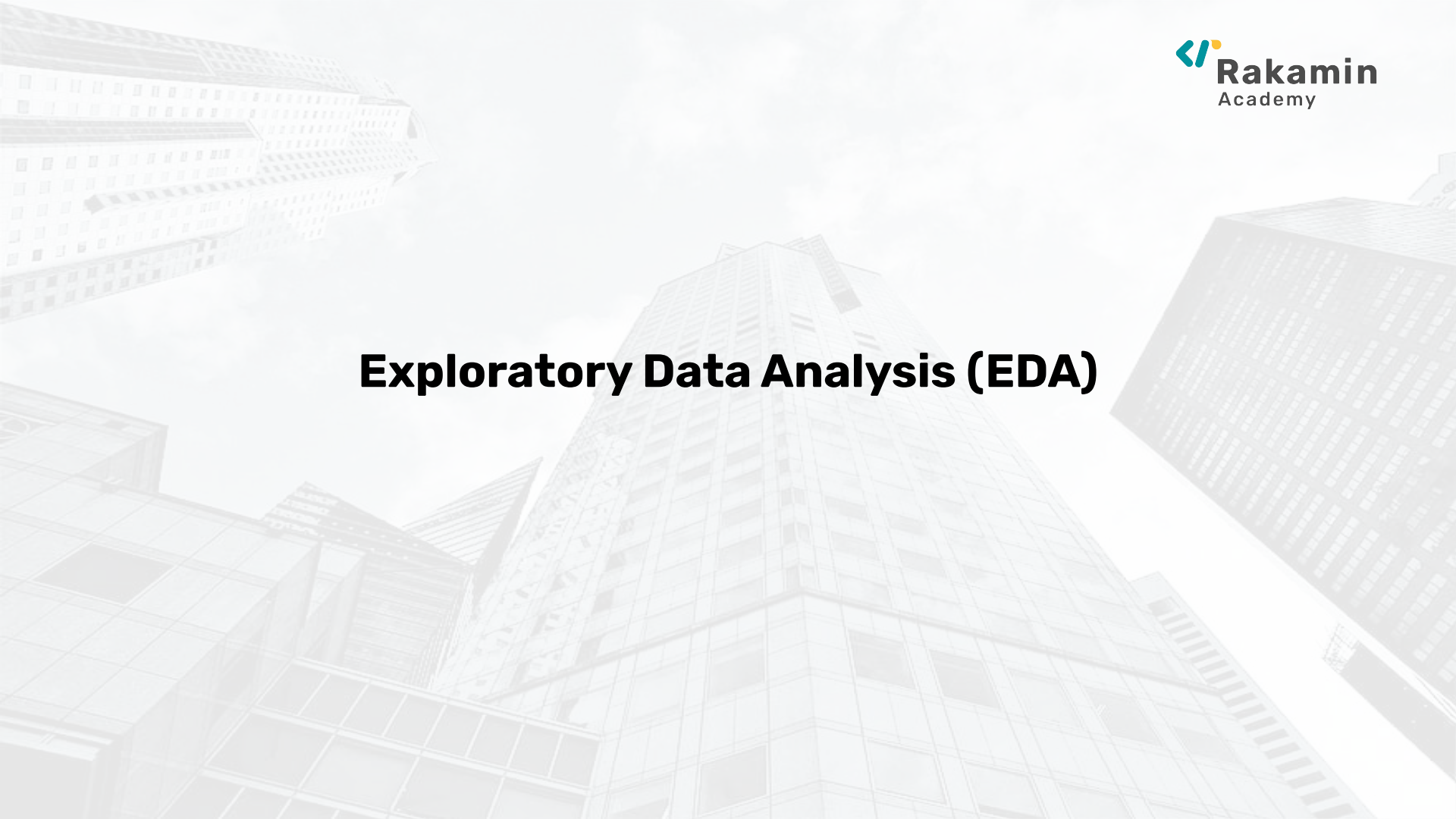
Case Study

Machine Learning- Clustering

Rakamin Academy

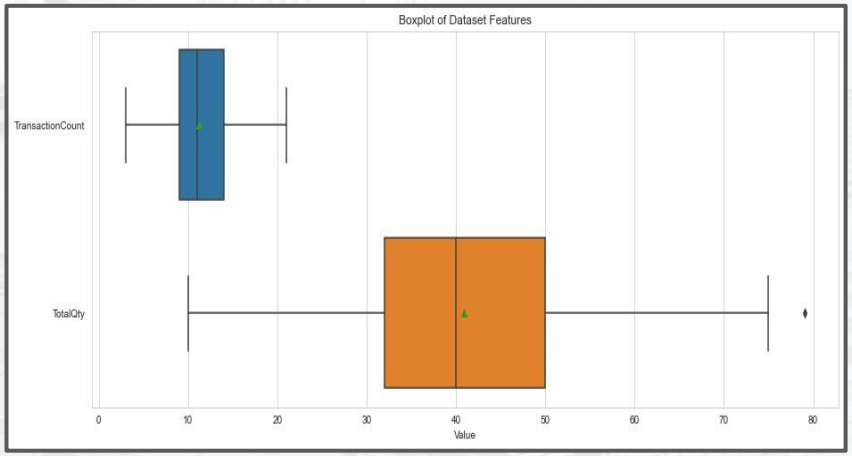
Dataset

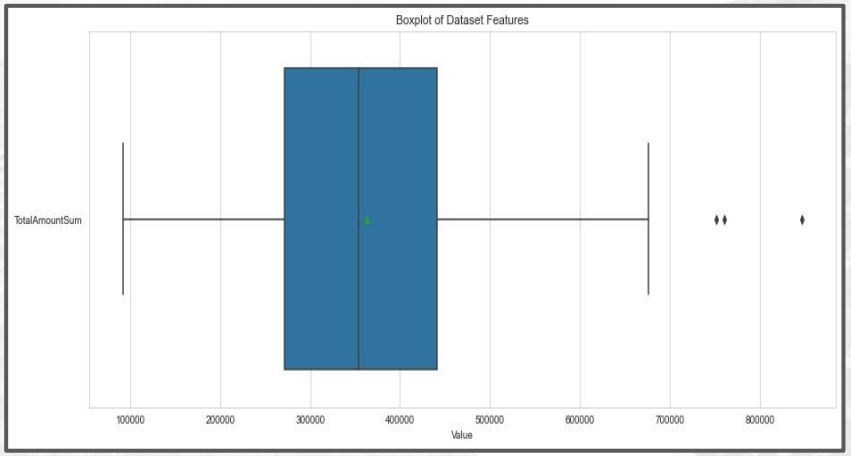
CustomerID	TransactionCount	TotalQty	TotalAmountSum
1	17	60	623300
10	14	50	478000
100	8	35	272400
101	14	44	439600
102	15	57	423300



Exploratory Data Analysis (EDA)

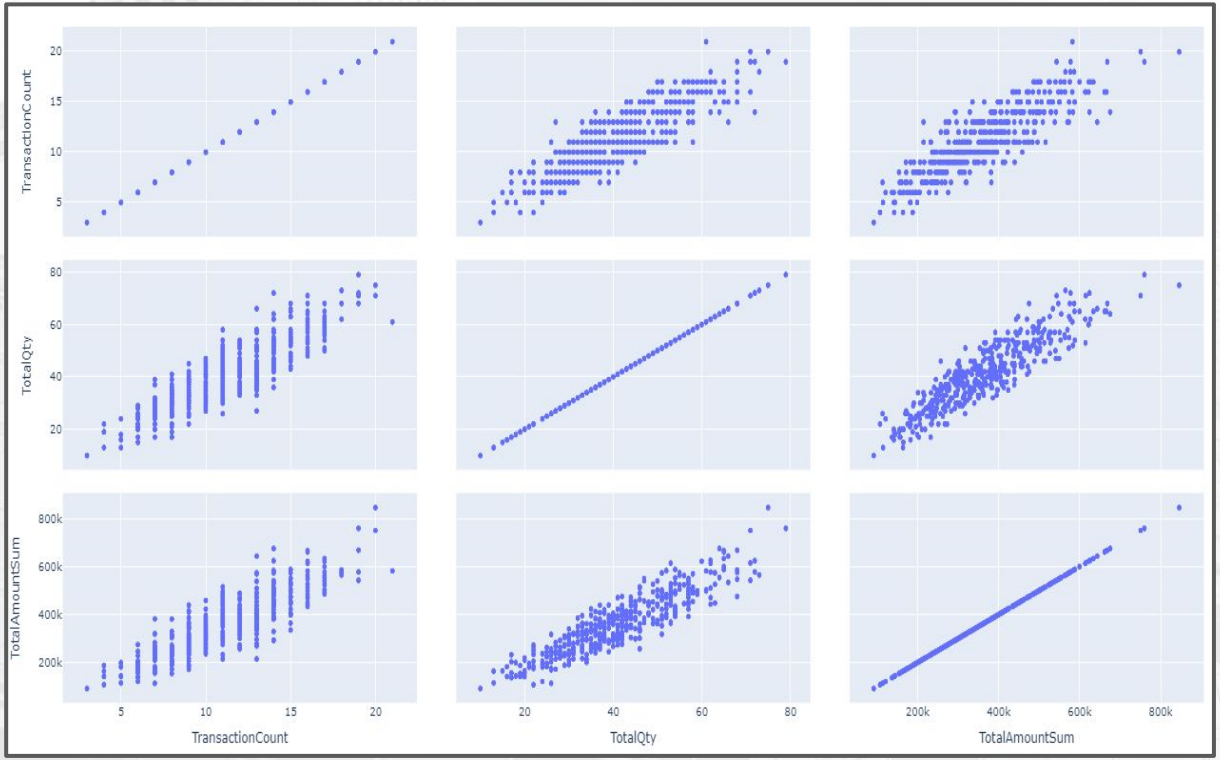
Drop Outlier





Dari kedua plot diatas dapat dilihat terdapat beberapa point yang valuenya outlier.
Oleh karena itu, nanti value-value tersebut akan didrop

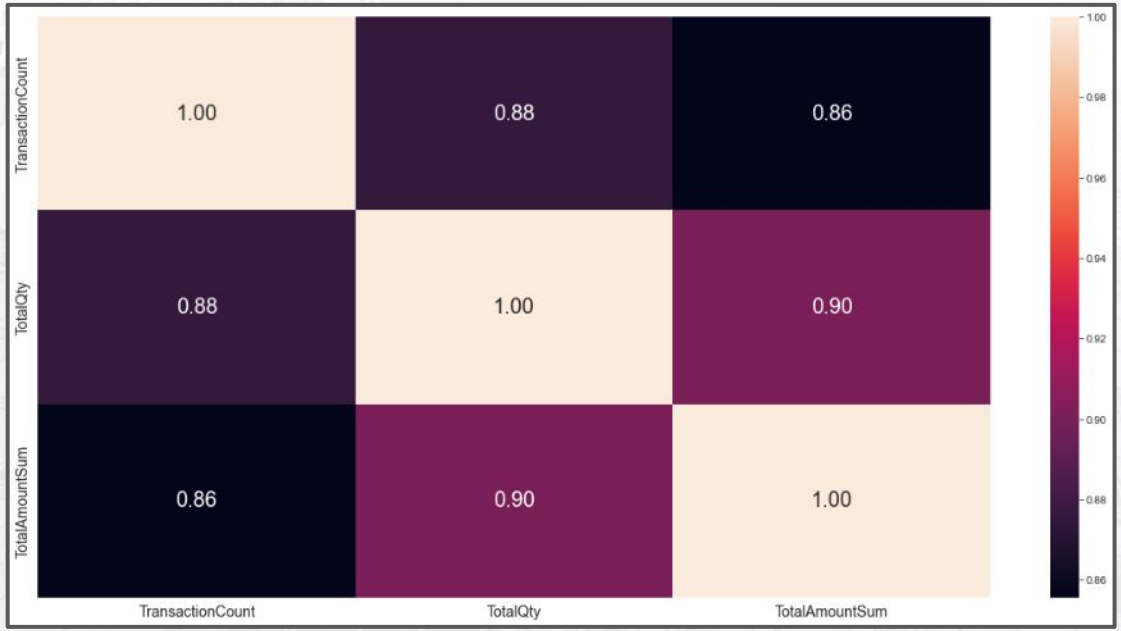
Data Distribution



Dari plot disamping, persebaran datanya mengarah ke kanan atas. Mengartikan bahwa data tersebut saling berkorelasi positif.

Jadi jika suatu feature valuenya naik, nanti feature-feature lainnya juga ikut menaik, dikarenakan berkolerasi positif.

Data Correlation



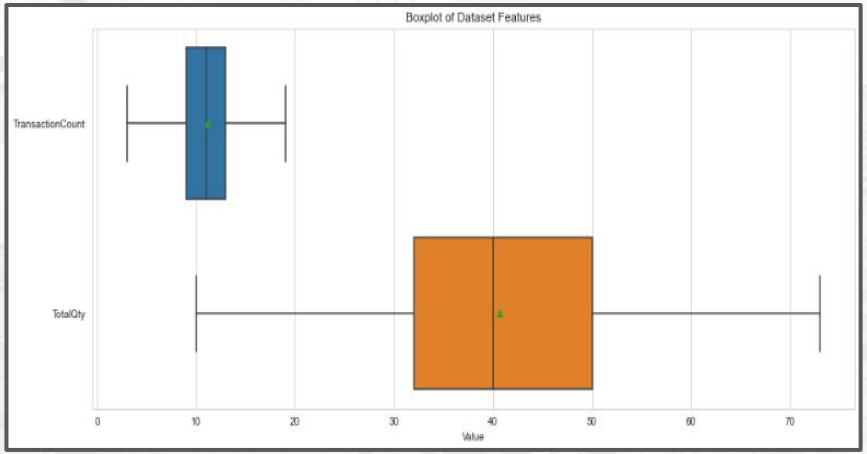
Berdasarkan slide sebelumnya, bahwa setiap featurenya berkolerasi positif.

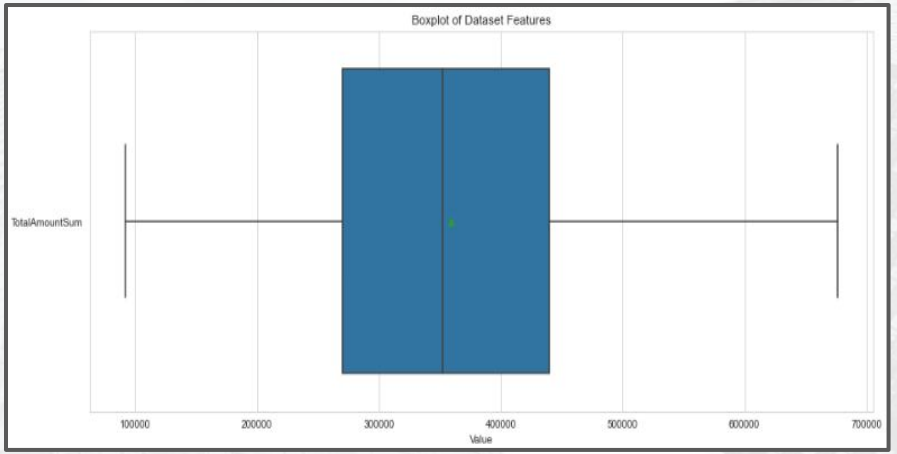
Ketika diplot kolerasi setiap featurenya menghasilkan value seperti disamping.

Dimana angkanya berada di angka 0.86 s/d 0.9.

Feature Engineering

Drop Outlier





Dari fase EDA (Exploratory Data Analysis) sebelumnya, akan ada membuang outlier.
Dari kedua plot diatas dapat dilihat bahwa data point yang outlier sudah didrop.

Min-Max Scaling

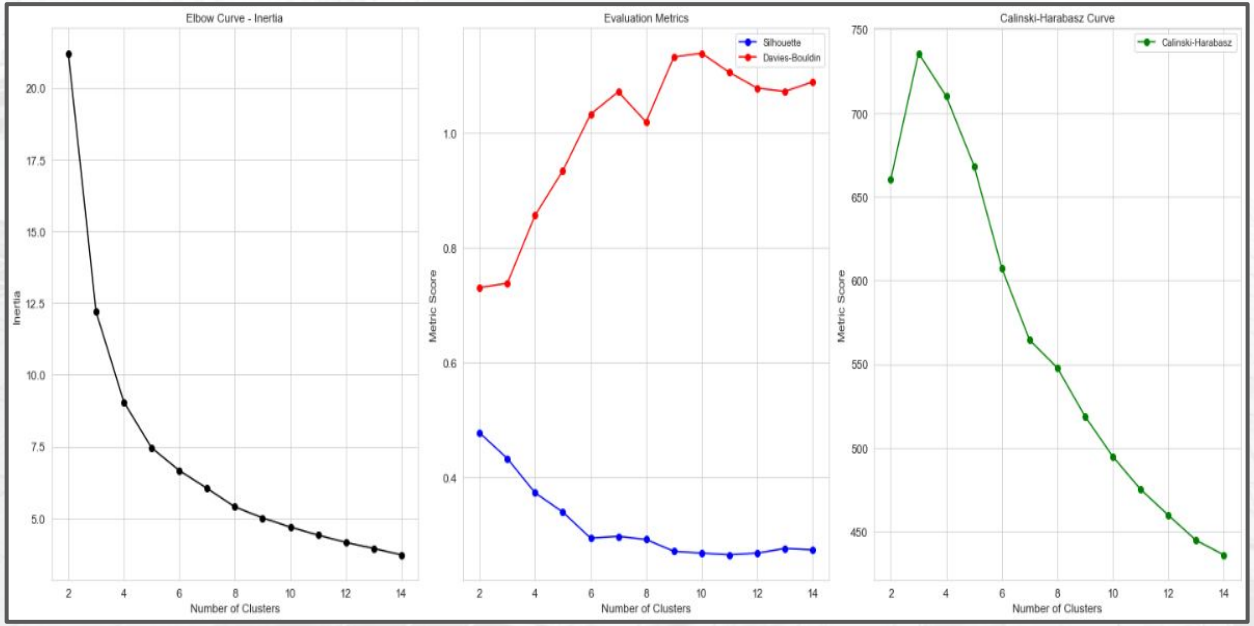
CustomerID	TransactionCount	TotalQty	TotalAmountSum
1	17	60	623300
10	14	50	478000
100	8	35	272400
101	14	44	439600
102	15	57	423300

Dari value dataframe disamping, dapat dilihat bahwa valuenya berada pada range yang berbeda-beda. Oleh karena itu, harus dilakukan min-max scaling, tujuannya supaya range valuenya berada pada range yang sama.

```
[[0.875, 0.7936507936507936, 0.9094333162129773],
 [0.6875, 0.6349206349206349, 0.6606745420304743],
 [0.3125, 0.39682539682539686, 0.30868002054442734],
 [0.6875, 0.5396825396825397, 0.5949323745933915],
 [0.75, 0.7460317460317459, 0.5670261941448382]]
```

Gambar disamping merupakan value setiap kolom setelah dilakukan min-max scaling. Dimana valuenya sudah pada rentang yang sama, yaitu 0 s/d 1 tanpa mengurangi makna dari datanya. Jadi hanya mengubah valuenya saja.

Clustering Metrics

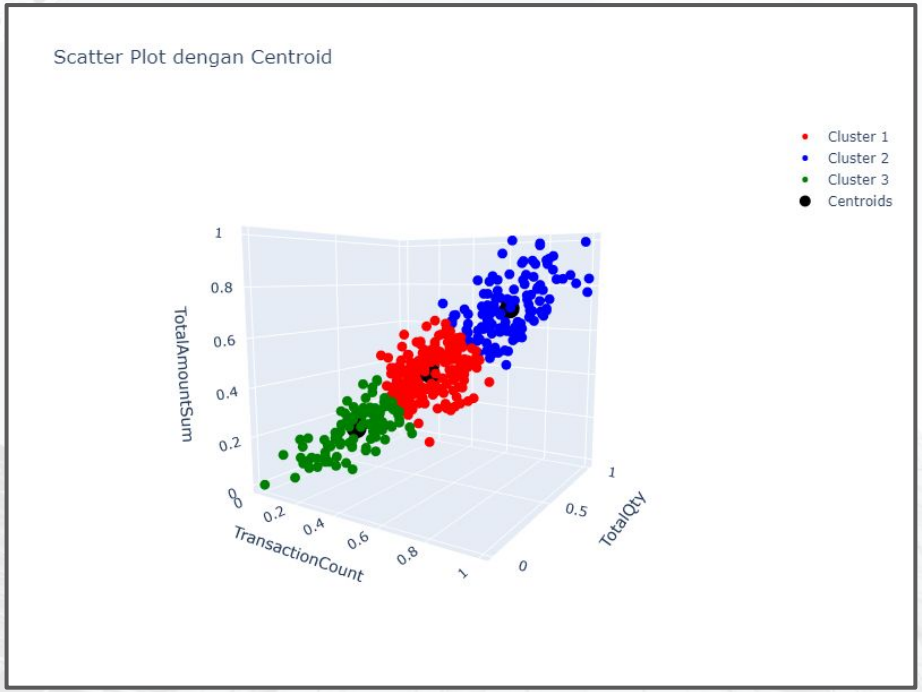


Saya menggunakan empat metrics, yaitu Elbow curve, Silhouette score, Davies-Bouldin (DB) score, CH-score.

1. Elbow method :
   Nilai k yang diambil ketika curvenya sudah patah dan mulai lambai.
2. Silhouette score :
   Semakin tinggi scorenya, semakin bagus clusternya.
3. DB score :
   Semalin rendah scorenya, semakin bagus clusternya.
4. CH (Calinki-Harabasz) score :
   Semakin tinggi valuenya, maka semakin bagus clusternya.

Dari keempat metrics ini, saya menggunakan k-clustering sebesar 3.

Scatter Plot Clustering



Gambar diatas merupakan scatter plot dengan jumlah k-clustering = 3.

Hasil Akhir

	TransactionCount	TotalQty	TotalAmountSum	Jumlah Customer
1	14.846774	55.887097	505401.612903	124
2	7.408696	26.034783	220050.434783	115
3	11.014706	39.637255	348654.901961	204

Tabel diatas merupakan rata-rata value dari TransactionCount, TotalQty, dan TotalAmountSum. Serta terdapat kolom jumlah customer.

Cluster dengan jumlah uang yang dikeluarkan terbanyak yaitu pada cluster 1. Sedangkan yang paling sedikit yaitu pada cluster .

```
~ ~ ~ Model Metrics Scores ~ ~ ~
Silhouette Score        : 0.43339306239644226
Davies-Bouldin Score    : 0.7387372361994299
Calinski-Harabasz Score : 735.6779809934534
```

Gambar diatas ini merupakan metrics score dari model yang dibuat.

Rakamin Academy

Case Study

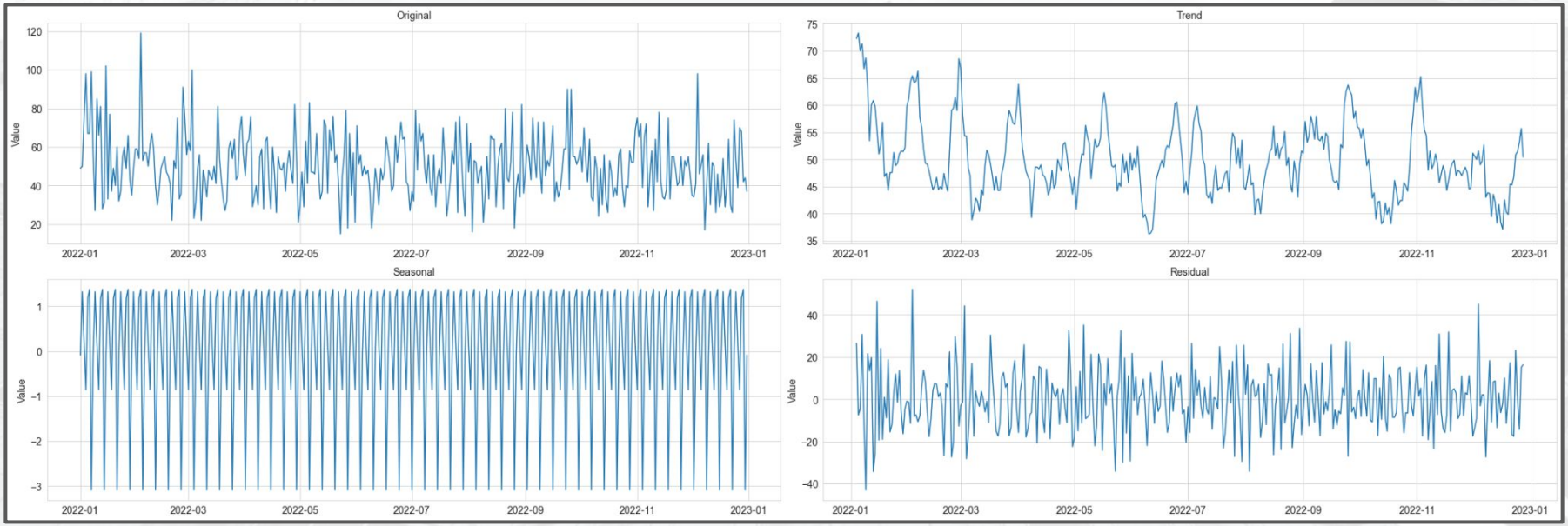
Machine Learning - Regression (Time Series)

Dataset yang Digunakan

Date	Qty
2022-01-01	49
2022-01-02	50
2022-01-03	76
2022-01-04	98
2022-01-05	67

Data disamping merupakan data yang akan dipakai untuk dibuat model ARIMA. Dimana kolom “Qty” tersebuh sudah dilakukan sebuah agregasi SUM dan di-groupby berdasarkan Date.

Dekomposisi Dataset



Uji Stasioner

Uji stationer yang saya gunakan, menggunakan dua metode :

1. Augmented Dickey – Fuller (ADF) Test
2. Kwiatkowski - Phillips - Schmidt - Shin (KPSS) Test

Uji Stasioner

ADF Test

	Keterangan	Nilai uji
0	Uji Statistik	-19.018783
1	p-value	0.0
2	Lags digunakan	0
3	Banyak Observasi	364
4	Hasil Uji	H0 ditolak
5	Kesimpulan	Data terindikasi Stasioner

Jika p-value < 0.05. Maka H0 ditolak dan berarti data stasioner.
Sebaliknya, jika p-value > 0.05. Maka H0 diterima dan artinya data non- stasioner

Keterangan ADF Test

Hipotesa Null (H0) : Data terindikasi non-stasioner.
Hipotesa Alternatif (H1) : Data terindikasi stasioner

Uji Stasioner

KPSS Test

	Keterangan	Nilai uji
0	Uji Statistik	0.425352
1	p-value	0.066227
2	Lags digunakan	5
3	Hasil Uji	H0 diterima
4	Kesimpulan	Data terindikasi Stasioner

Jika p-value < 0.05. Maka H0 ditolak dan berarti data non-stasioner.
Sebaliknya, jika p-value > 0.05. Maka H0 diterima dan artinya data stasioner

Keterangan KPSS Test

Hipotesa Null (H0) : Trend data terindikasi stasioner.
Hipotesa Alternatif (H1) : Data terindikasi non-stasioner

Uji Stasioner

ADF Test

	Keterangan	Nilai uji
0	Uji Statistik	-19.018783
1	p-value	0.0
2	Lags digunakan	0
3	Banyak Observasi	364
4	Hasil Uji	H0 ditolak
5	Kesimpulan	Data terindikasi Stasioner

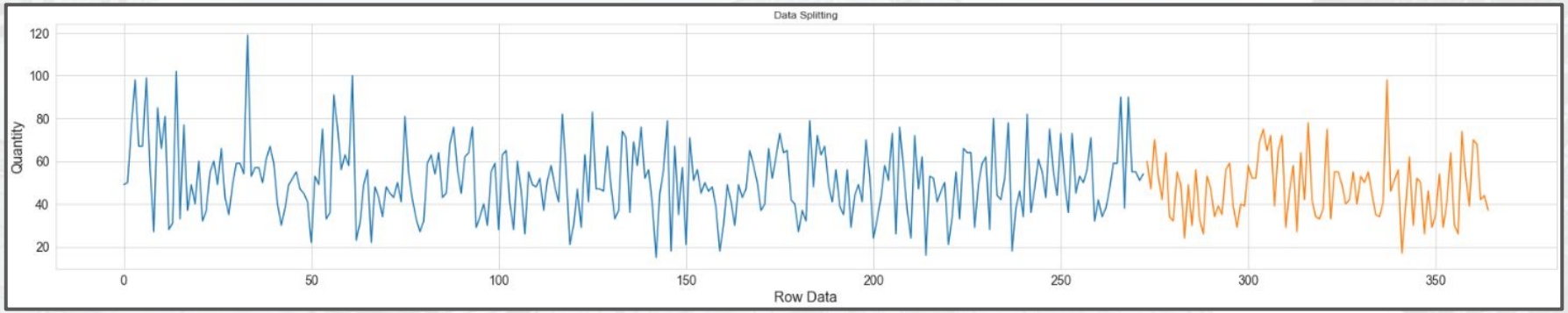
KPSS Test

	Keterangan	Nilai uji
0	Uji Statistik	0.425352
1	p-value	0.066227
2	Lags digunakan	5
3	Hasil Uji	H0 diterima
4	Kesimpulan	Data terindikasi Stasioner

Dari kedua test ini didapatkan kesimpulan bahwa Data terindikasi stasioner. Dikarenakan pada kedua uji test ini mengindikasikan bahwa datanya stasioner.

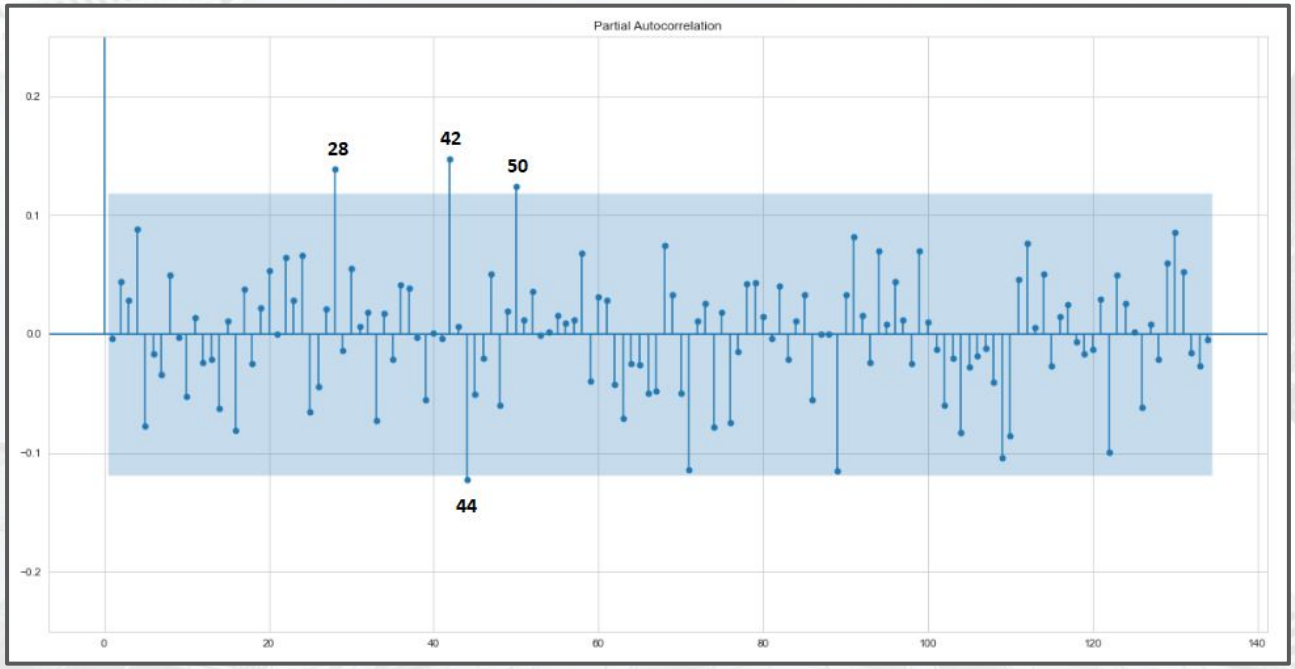
Dengan informasi ini, nanti untuk ordo d pada model ARIMA, nanti nilainya dijadikan 0. Karena tidak perlu dilakukan differencing kembali pada datanya.

Data Splitting



Data train saya gunakan 70% dan sisanya (30%) merupakan data testnya.
Datanya nanti digunakan untuk penentuan ordo p dan q.

Ordo p - PACF (Partial Autocorrelation Function)

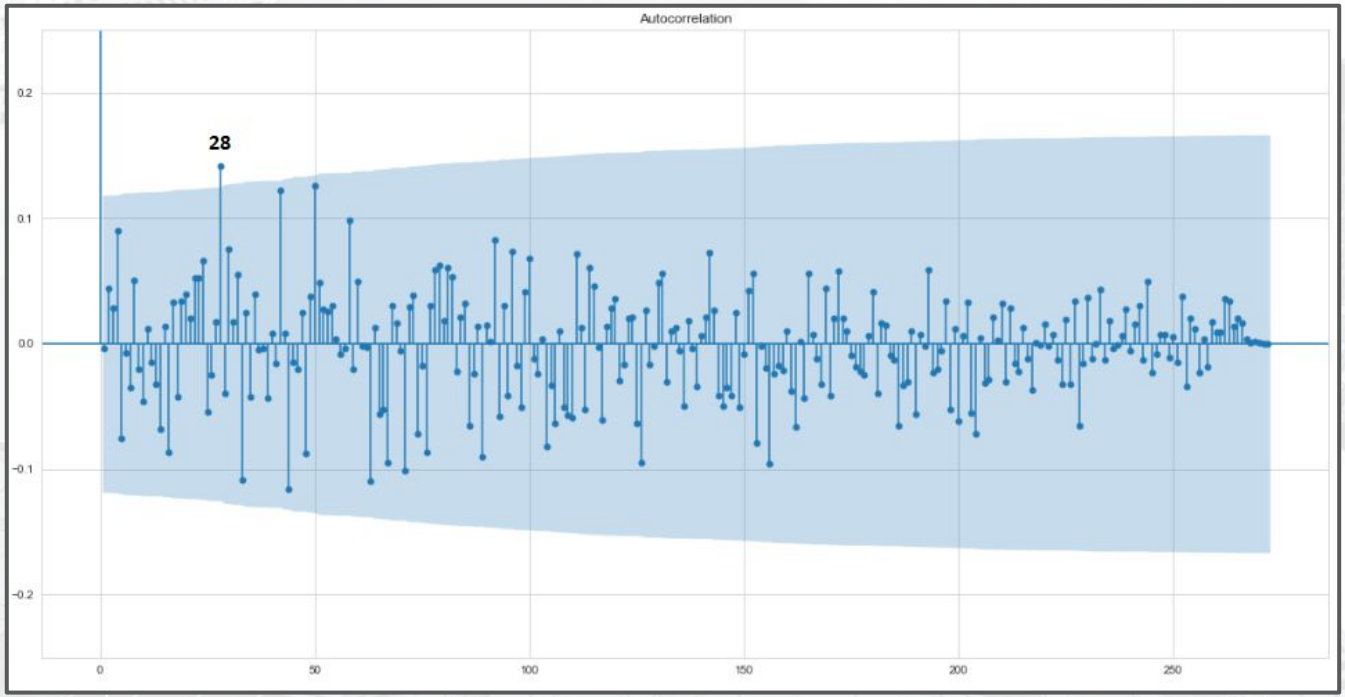


Dari plot PACF diatas, terdapat empat lag yang diluar bluish area (area yang berwarna biru) yaitu lag 28, 42, 44, dan 50.

Ordo d

Ordo d saya tentukan menjadi 0. Dikarenakan datanya sudah stasioner, jadi tidak perlu dilakukan differencing dengan ordo d ini.

Ordo q - ACF (Autocorrelation Function)



Dari plot ACF diatas, terdapat satu lag yang diluar bluish area (area yang berwarna biru) yaitu lag 28.

Auto ARIMA

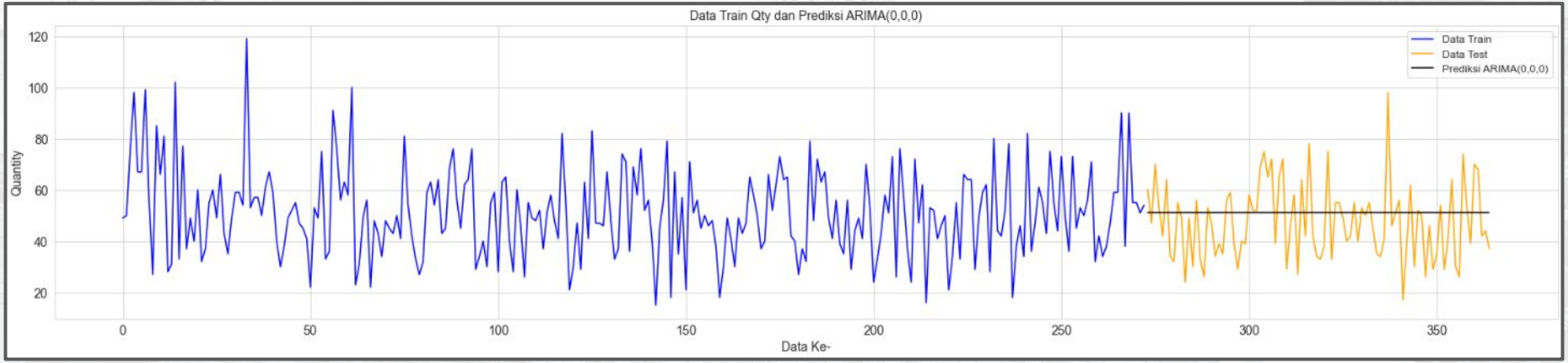
```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(0,0,0)[0]             : AIC=2342.651, Time=0.19 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=2952.911, Time=0.01 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=2508.940, Time=0.02 sec
 ARIMA(0,0,1)(0,0,0)[0]             : AIC=2776.756, Time=0.06 sec
 ARIMA(1,0,2)(0,0,0)[0]             : AIC=2341.934, Time=0.16 sec
 ARIMA(0,0,2)(0,0,0)[0]             : AIC=2676.272, Time=0.08 sec
 ARIMA(1,0,1)(0,0,0)[0]             : AIC=2340.029, Time=0.09 sec
 ARIMA(2,0,1)(0,0,0)[0]             : AIC=2341.929, Time=0.18 sec
 ARIMA(2,0,0)(0,0,0)[0]             : AIC=2433.732, Time=0.04 sec
 ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=2334.527, Time=0.08 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=2332.524, Time=0.05 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=2330.528, Time=0.01 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=2332.524, Time=0.04 sec

Best model:  ARIMA(0,0,0)(0,0,0)[0] intercept
Total fit time: 1.038 seconds
```

Dari hasil Auto ARIMA untuk menentukan ordo (p,d,q). Disimpulkan bahwa ordo terbaik yaitu (0,0,0) dikarenakan nilai AIC terkecil yaitu sebesar 2330.528.

Namun berdasarkan plot PACF (ordo p) dan ACF (ordo q) sebelumnya.
Untuk ordo p terdapat empat lag yaitu 28, 42, 44, dan 50.
Lalu ordo q terdapat satu lag yaitu 28.

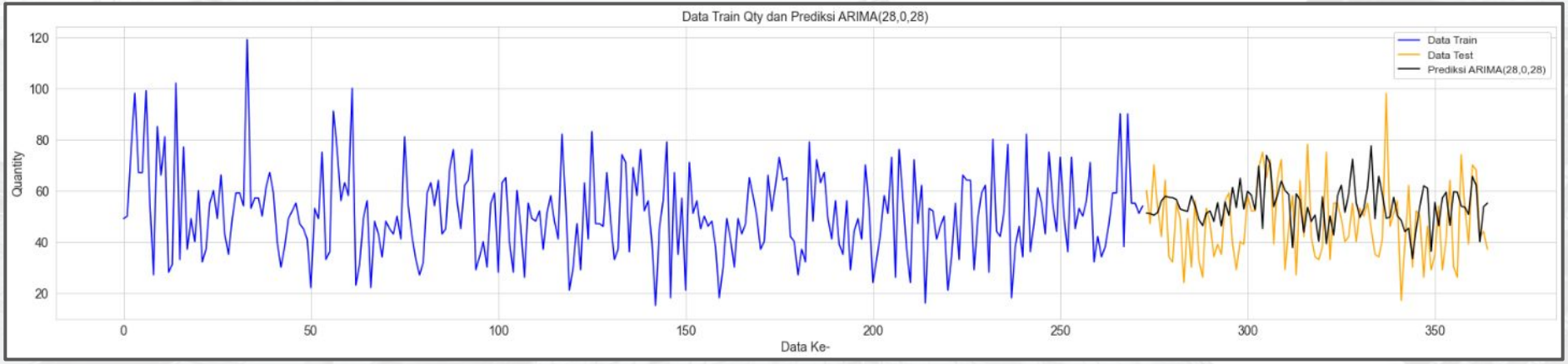
Maka saya mencoba kemungkinan-kemungkinan ordo p,d,q dengan kombinasi angka 28, 42, 44, dan 50.
Lalu saya mencari nilai MSE dan MAE menggunakan cross-validation supaya dapat membandingkan dengan model-model lainnya.

Ordo (0, 0, 0)



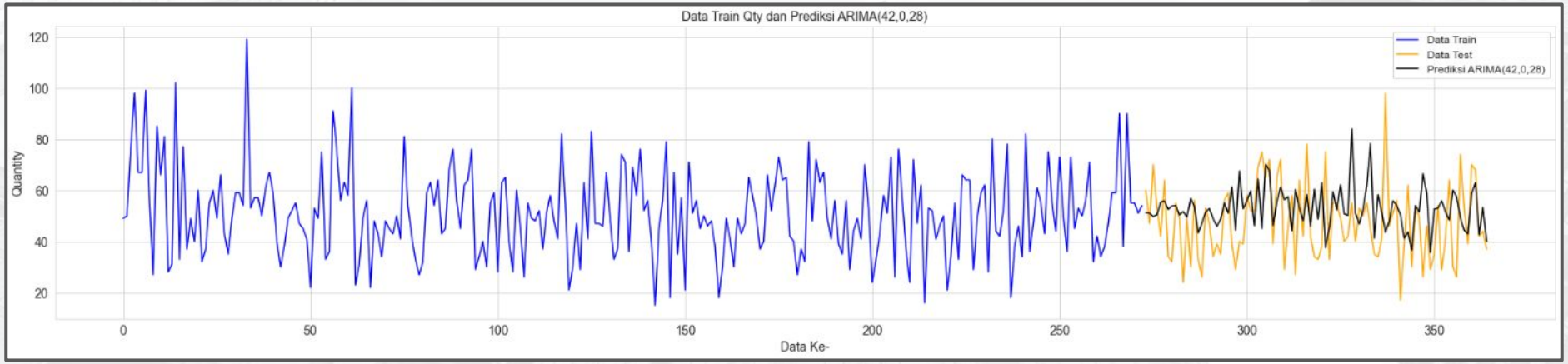

```
Average MSE: 277.1196265119027
Average MAE: 13.43943779632897
```

Ordo (28, 0, 28)



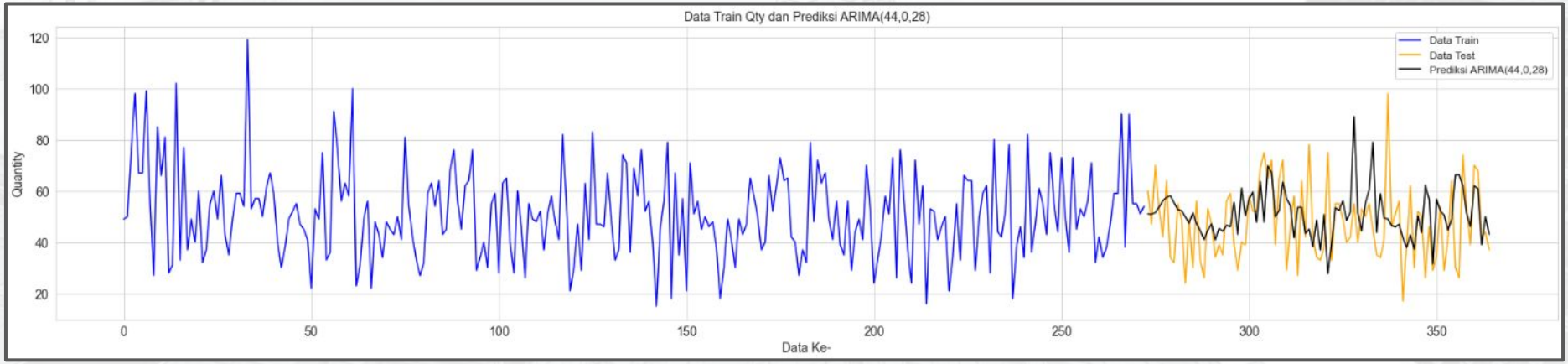

```
Average MSE: 507.5467529390553
Average MAE: 17.20640598321131
```

Ordo (42, 0, 28)



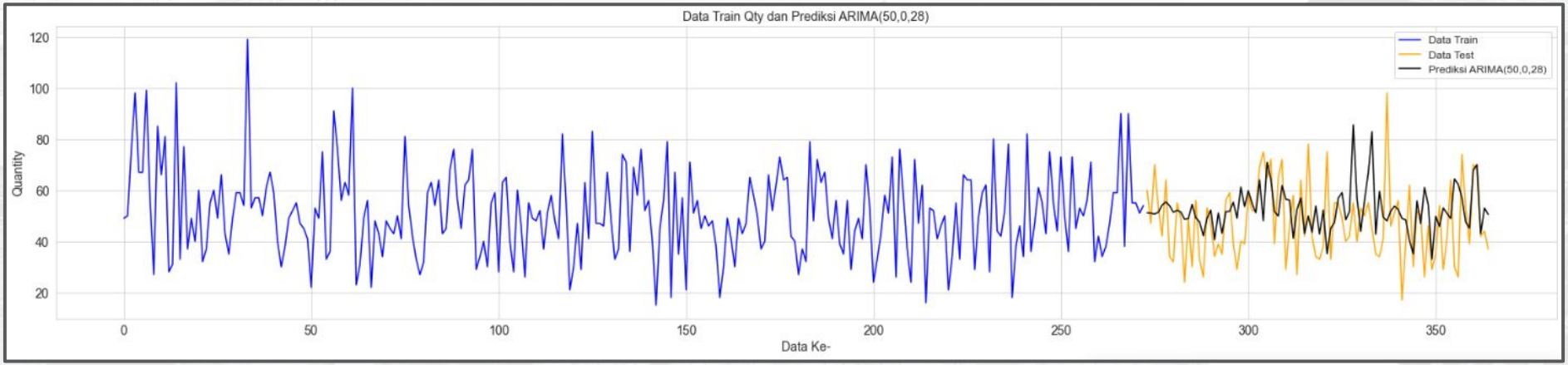

```
Average MSE: 552.7357265162906
Average MAE: 17.46957313266143
```

Ordo (44, 0, 28)




```
Average MSE: 473.0635526322021
Average MAE: 16.532379933269432
```

Ordo (50, 0, 28)




```
Average MSE: 610.9998536076489
Average MAE: 18.248802427565835
```

Conclusion

Dari seluruh ordo (p,d,q) yang saya coba.
Saya memutuskan untuk menggunakan ordo (0,0,0) atau (28, 0, 28).
Dikarenakan kedua kombinasi ordo (p,d,q) ini memiliki nilai MSE dan MAE yang terkecil.

Rakamin Academy

Case Study

Tableau Public Dashboard

[Final Task - Kalbe | Tableau Public]

Challenge

Query 1 : Berapa rata-rata umur customer jika dilihat dari marital statusnya ?

```
SELECT ROUND(AVG(c.age), 2) AS Avg_age, c."Marital
Status"
FROM "Transaction" t
INNER JOIN customer c
ON c.customerid = t.customerid
GROUP BY 2
```

Output :

	avg_age	Marital Status
1	31.41	
2	43.37	Married
3	29.69	Single

Query 2 : Berapa rata-rata umur customer jika dilihat dari gender nya ?

```
SELECT ROUND(AVG(c.age), 2) AS Avg_age, c.gender
FROM "Transaction" t
INNER JOIN customer c
ON c.customerid = t.customerid
GROUP BY 2
```

Output :

	avg_age	gender
1	40.39	0
2	39.54	1

Challenge

Query 3 : Tentukan nama store dengan total quantity terbanyak!

```
SELECT s.storename, SUM(t.qty) AS quantity
FROM store s
INNER JOIN "Transaction" t
ON s.storeid = t.storeid
GROUP BY 1
ORDER BY quantity DESC
LIMIT 1
```

Output :

	storename	quantity
1	Lingga	2,777

Query 4 : Tentukan nama produk terlaris dengan total amount terbanyak!

```
SELECT p."Product Name", SUM(t.totalamount) AS total_amount
FROM product p
INNER JOIN "Transaction" t
ON p.productid = t.productid
GROUP BY 1
ORDER BY total_amount DESC
LIMIT 1
```

Output :

	Product Name	total_amount
1	Cheese Stick	27,615,000

Rakamin Academy

Insert Your Link Github Here

Github Profile : TheOX7 (Marselius Agus Dhion) (github.com)
Github Repository : TheOX7/Final-Task-Kalbe (github.com)

Rakamin Academy

Video Presentation Here

Link Video Presentation

Thank You